iSTART

April 2026

iSTART iReport

芯测科技电子报第 15 期

芯测科技 IEEE 1838 解决方案到位
抢滩 AI Chiplet 高阶封装商机

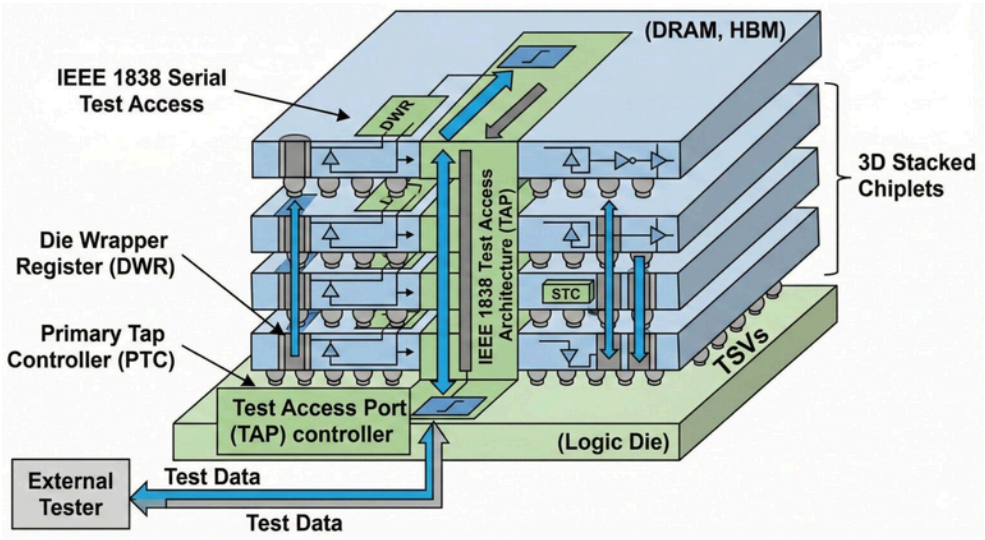
AI 计算需求快速攀升，半导体产业正全面迈向以 3D-IC 与 Chiplet 为核心的先进封装时代。高带宽内存 (HBM) 与逻辑芯片的垂直堆栈，已成为支撑 AI 加速器效能与能效的关键技术路线。其中以 CoWoS 为代表的 2.5D/3D 封装架构，虽能大幅提升系统整体效能，却也突显了测试、修复与良率控管上前所未有的挑战。

专注于内存测试与修复技术的芯测科技 (iSTART-TEK) ，宣布其 IEEE 1838 3D-IC 测试解决方案已进入成熟应用阶段，是目前少数能完整对应 3D-IC 测试标准、并实际落实于先进封装项目的 EDA 供货商之一，提前卡位 AI Chiplet 与 HBM 高阶封装商机。

率先布局 IEEE 1838，建立 3D-IC 测试门坎

IEEE 1838 为专为 3D-IC 与异质整合架构制定的测试标准，涵盖 die-to-die 测试存取、测试信道重组、以及堆栈后的可测试性设计。然而由于技术门坎高、整合复杂度大，目前市场上能成熟支持 IEEE 1838 的 EDA 工具选项相当有限。

芯测科技早在 3D-IC 尚未全面商用之前，便已投入相关架构与测试流程的研发，将 IEEE 1838 纳入其内存测试与修复平台中，并完成与既有 BIST、BISR 机制的整合。此率先布局不仅让芯测科技在 3D-IC 测试领域建立技术护城河，也大幅提高后进者的进入门坎。



深度对应异质整合，满足 HBM 与逻辑芯片堆栈需求

AI 芯片的效能瓶颈，已不再仅限于计算单元，还高度取决于 HBM 与逻辑芯片之间的数据传输效率。为此，HBM 与 GPU、AI 加速器、客制化逻辑芯片的垂直整合，成为主流设计方向，也让测试架构必须同时理解「内存」与「逻辑」的堆栈关系。

芯测科技的 IEEE 1838 解决方案，即是针对 HBM 与逻辑芯片的异质整合情境所设计，可支持多颗 die 堆栈后的测试路径规划、测试存取管理，以及内存层级的故障定位与修复策略。透过此架构，客户得以在 CoWoS 等先进封装流程中，保有对 HBM 与 SRAM 的可测试性与可修复性，确保 AI 系统的效能与可靠度。

降低 CoWoS 报废风险，放大先进封装投资回报

在 CoWoS 与 3D-IC 架构下，一颗整合 HBM 与高阶逻辑芯片的成品，往往代表极高的制造成本。一旦在封装后才发现内存或互连缺陷，若无有效的测试与修复机制，报废损失将相当可观。

芯测科技透过 IEEE 1838 与其内存修复技术的整合，协助客户在多 die 堆栈完成后，仍能进行精准的故障诊断与修复，大幅提升最终良率。这不仅直接降低先进封装的报废成本，也让客户在导入 CoWoS 与 AI Chiplet 架构时，能有效放大整体投资报酬率 (ROI) 。

芯测全流程 eFlash BIST IP 服务
把关 AI 服务器关键组件质量

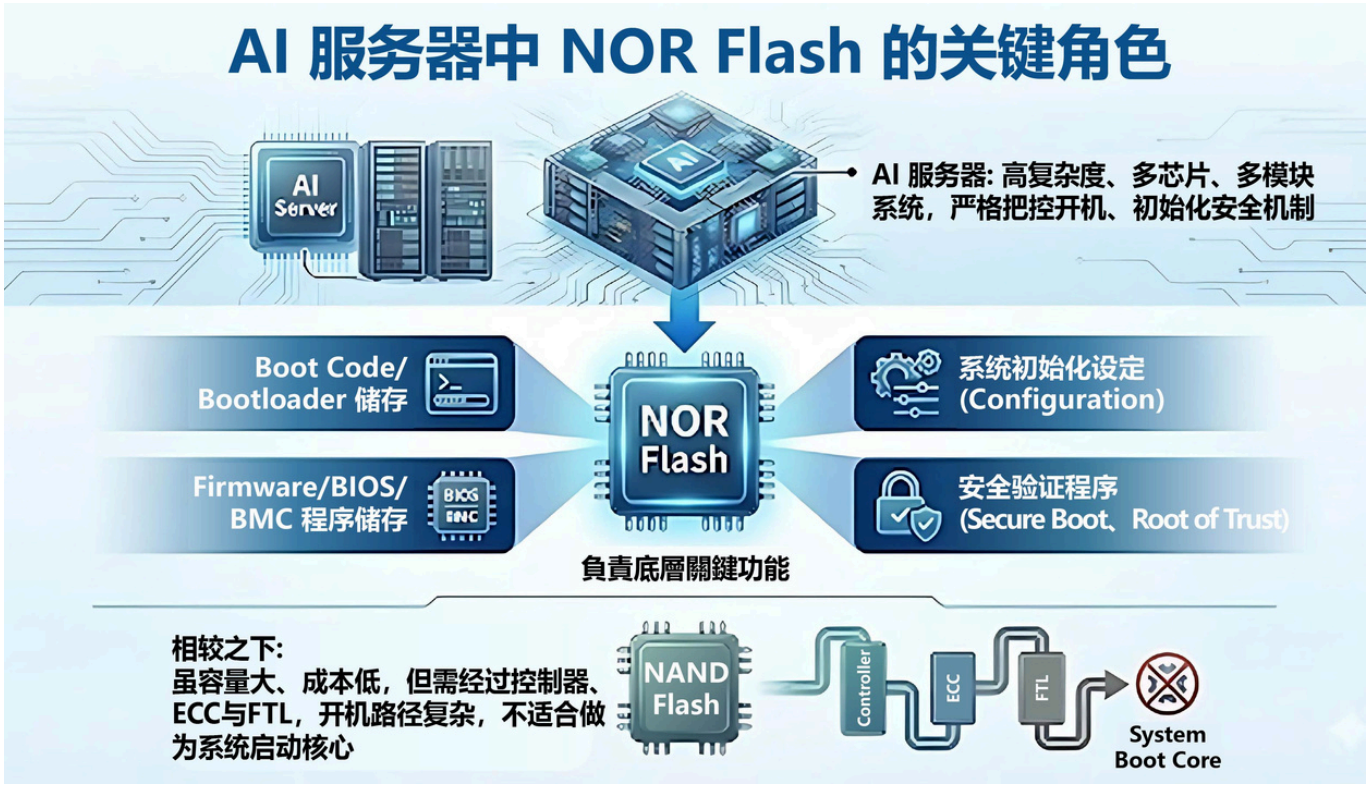
生成式 AI 与大型语言模型 (LLM) 爆发性成长，全球数据中心对于 AI 服务器的效能要求已达前所未有的高度。然而在追求算力的同时，「系统稳定性」与「数据可靠性」成为了隐形的技术天花板。内存测试与修复领导厂商芯测科技分析，AI 服务器中的关键组件 NOR Flash 虽看似微小，却是启动与安全的核心。

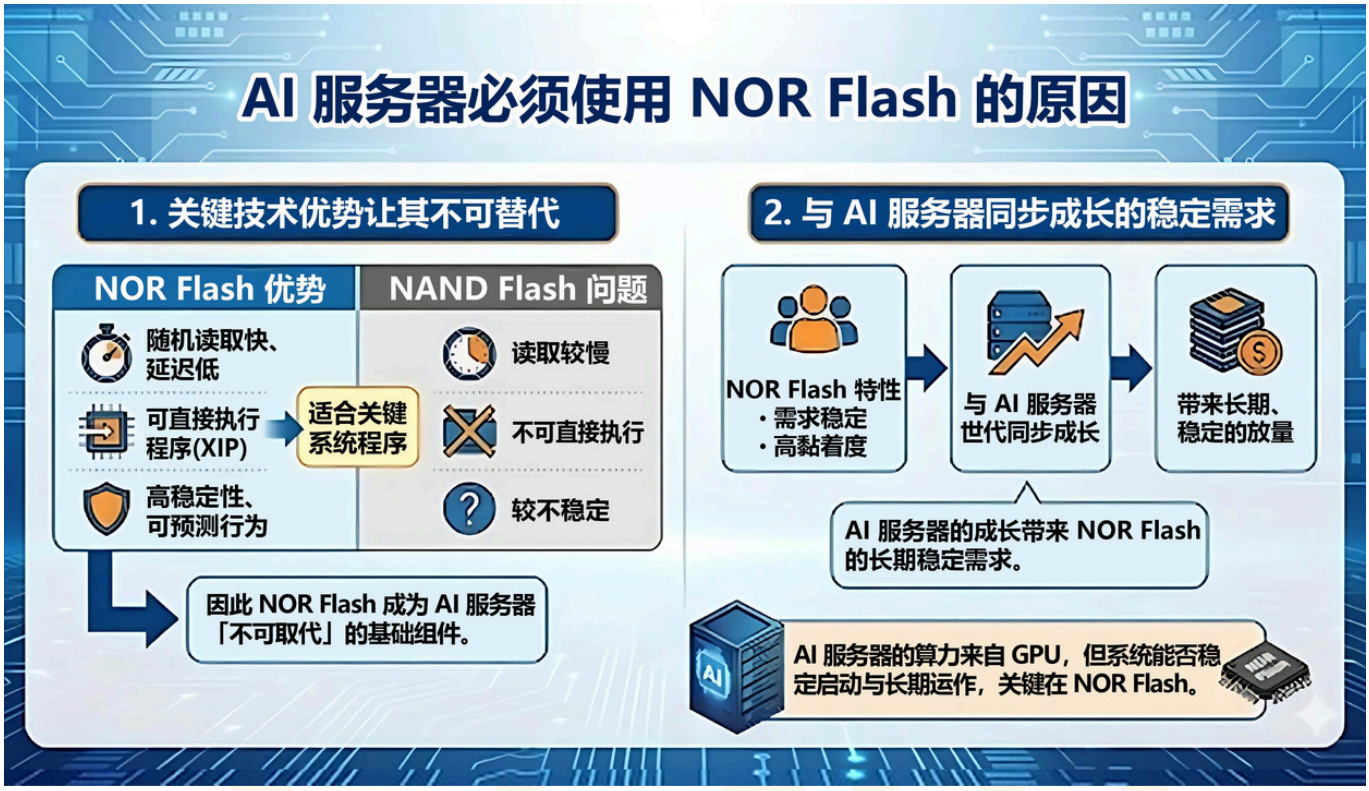
AI 服务器的核心痛点—— NOR Flash

在 AI 服务器的架构中，NOR Flash 负责存储 Boot Code 与 Bootloader，并具备低功耗与快速读取功能。随着系统复杂化，任何细微的内存故障都可能导致整台服务器宕机，造成巨大的算力损失。

在多数 AI 服务器架构中，NOR Flash 是启动与安全机制中高度关键的组件。针对此趋势，芯测提供的 **eFlash (NOR Flash) BIST IP** 正是为了解决以下设计痛点：

- **高检验标准：** 沿袭车用电子芯片的严苛检验标准，确保在高速计算的热循环环境下依然稳定。
- **生命周期监控：** 透过专利测试算法，于芯片生命周期不同阶段进行健康状态检测。

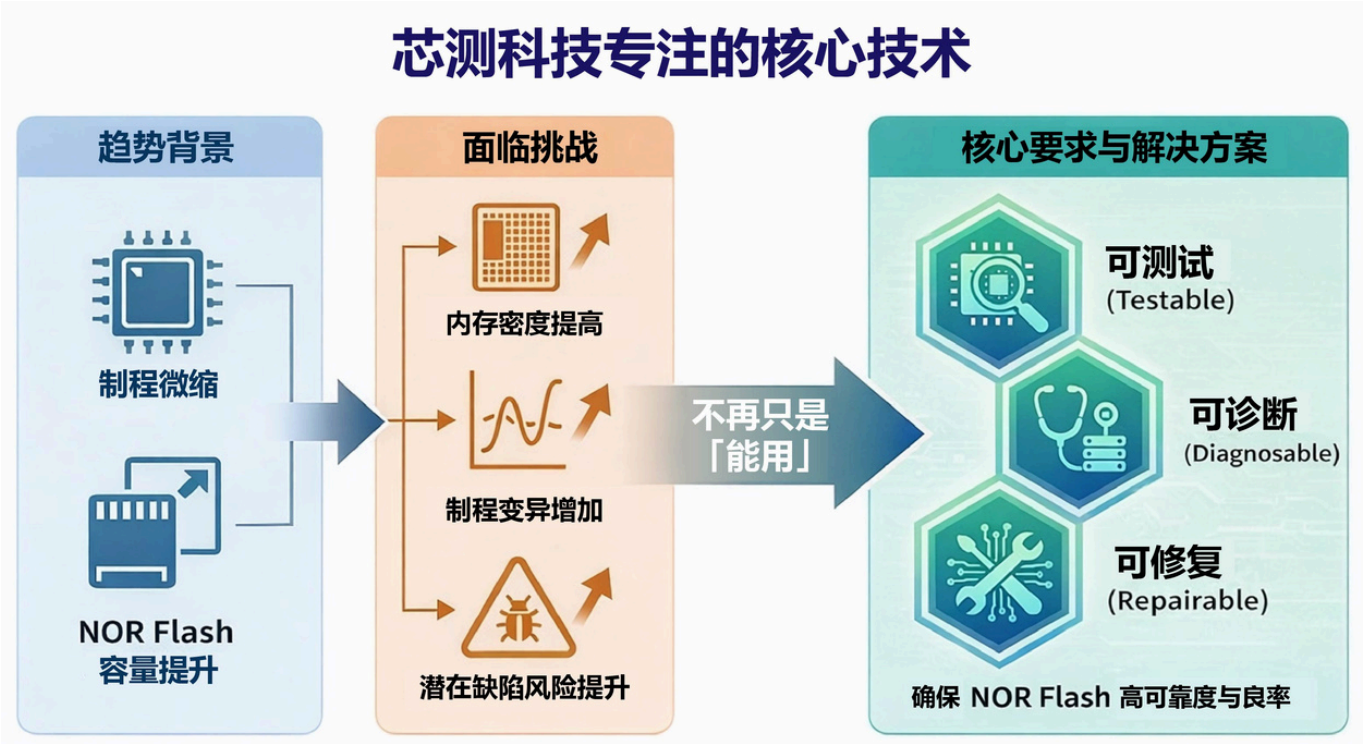




从 BIST 到 BISR 的完整技术链

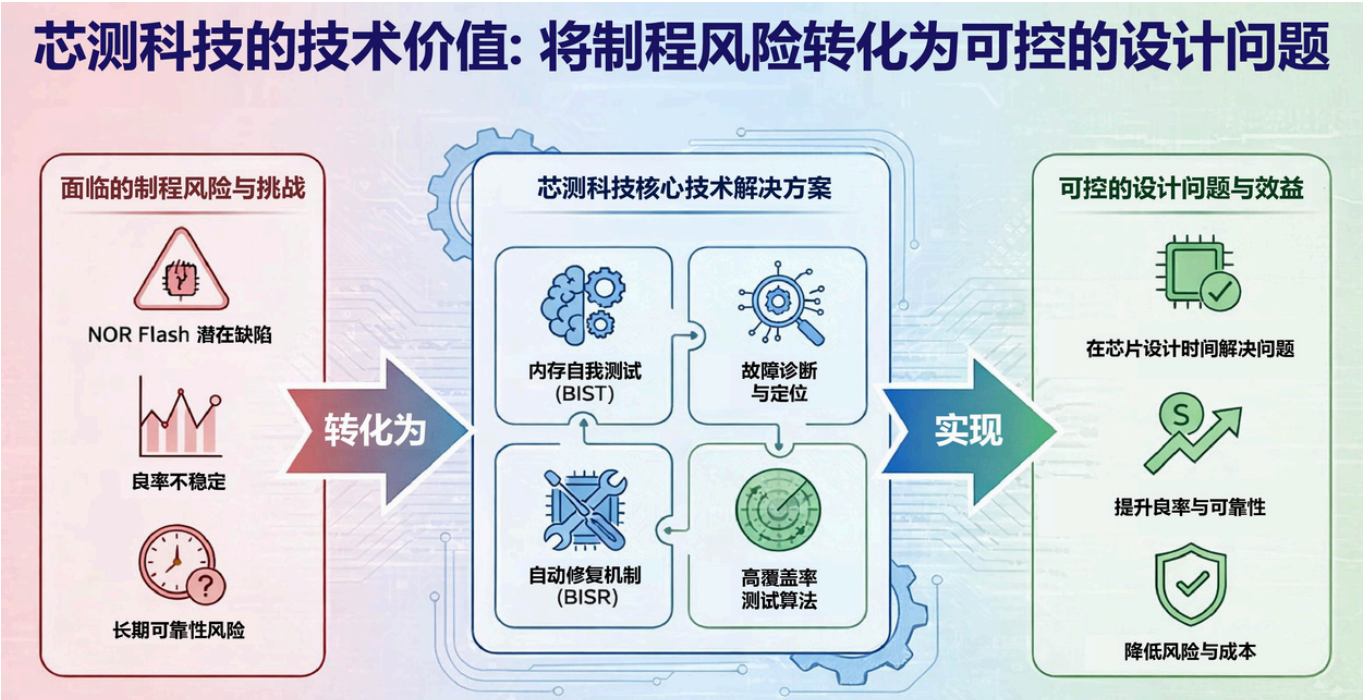
芯测科技的优势在于将复杂的测试流程转化为可控的设计效益。其核心技术解决方案包含四大模块：

- 内存自我测试 (BIST)： 提供极高的故障检出率。
- 故障诊断与定位： 缩短工程师除错时间，加速产品上市。
- 自动修复机制 (BISR)： 透过硬件修复大幅提升芯片良率，这对昂贵的 AI 芯片尤为重要。
- 高覆盖率测试算法： 针对先进制程 (如 eFlash 节点) 提供精准的测试覆盖。



产业领先地位：从车规级实绩到全方位设计服务

芯测科技之所以能快速切入 AI 供应链，得益于其长期在高安全关键领域的累积，并具备成功协助客户完成 55nm eFlash 制程设计的实战经验，包括从前端 IP 评估、BIST 设计，到后端布局绕线 (APR) 与 Turnkey 流片的全流程服务。这种全方位的交付能力，让芯片设计公司能专注于 AI 架构创新，而将复杂的测试工作与质量门坎交由芯测把关。



NPU 热潮下 Edge AI 正在重新定义 AI 计算架构

随着 AI 应用从云端走向终端设备，NPU (Neural Processing Unit) 成为近期产业讨论的焦点。相较于传统 CPU 与 GPU，NPU 以更低功耗与更高效率，让 AI 模型能直接在边缘端实时运行，不仅降低延迟，也减少对云端的依赖，同时强化数据隐私。

在 Edge AI 发展早期即深耕此领域的耐能智慧 (Kneron) ，长期专注于边缘 AI SoC 与 NPU 架构设计，透过「AI 芯片 × 边缘计算 × 图像算法」的整合，推动智能物联、自动驾驶与智能安防等应用落地。

芯测科技与耐能智慧有密切的合作经验。在耐能智能的技术交流与实务经验驱动下，芯测科技持续精进于内存测试与修复解决方案，协助先进 AI SoC 在量产阶段兼顾效能、良率与可靠度。

AI 推论新战局：SRAM 成为效能关键，如何抢占先机？

辉达 (NVIDIA) GTC 大会再次点燃 AI 战火！据媒体报导，辉达预计推出 V4 版本的 LPU，透过台积电 N3P 制程搭配 CoWoS-R 来打造，未来搭配 GPU、NVLink (高速互连) 整合为完整系统解决方案。

业者分析，随生成式 AI 由训练走向随着 AI 推论需求爆炸式成长，传统内存带宽已面临瓶颈，访问速度更快的 SRAM 正成为提升计算效能的「第一排」关键武器。当台积电与全球 IC 设计大厂纷纷强化 SRAM 布局时，如何确保芯片中庞大 SRAM 阵容的可靠性与良率，就是胜出的核心。

在此浪潮下，芯测科技所提供最全面的内存测试与修复解决方案，将可让企业夺得领先优势：

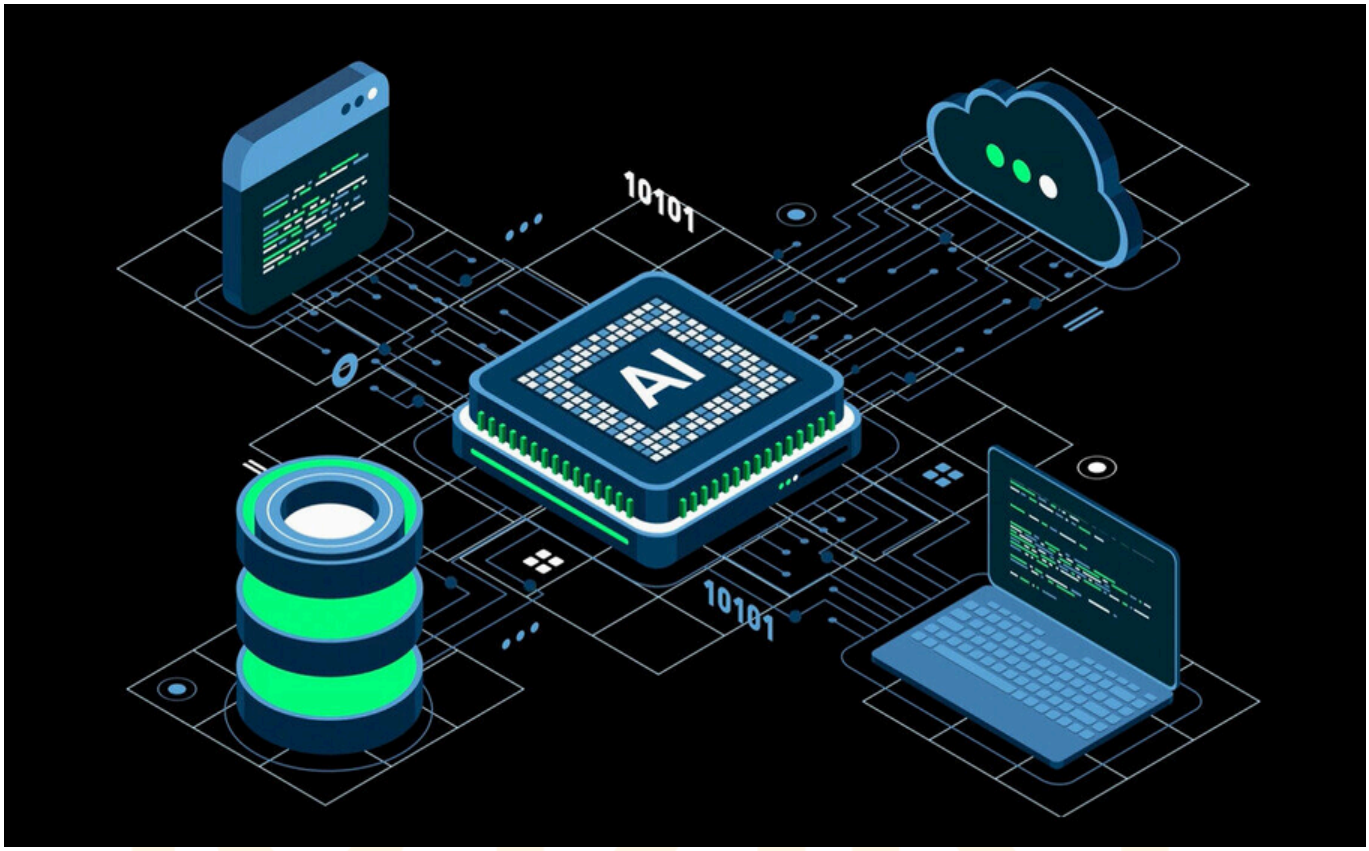
自动化开发平台：缩短内存测试电路设计时间，加速产品上市。

高效率修复技术：提升 SRAM 良率，将昂贵的先进制程成本降到最低。

应对复杂架构：ISO26262 TCL1车规等级技术，可轻松因应 AI 芯片高规格的内存需求。

在 AI 芯片追求极致的道路上，芯测始终是最可靠的伙伴。

SRAM Repair 正成为 AI 浪潮下的关键成长动能



从先进制程授权、PUF 导入 AI 架构，到云端 AI 与高效能计算需求扩张，产业资金与技术焦点正快速向AI 相关应用集中。从近期的市场讯号可明显看出：AI Data Center Processor、HBM高带宽内存、以及各类 AI 加速芯片，正在推升对嵌入式内存可靠度的要求。

先进制程 (3nm 及以下) 中，SRAM bit cell 尺寸持续缩小，制程变异、随机缺陷与老化效应更加敏感。在 AI 推论与训练架构中，大量的 on-chip SRAM、cache 与 buffer 被用于降低延迟与数据搬移成本，使得单颗 SoC 的 SRAM 容量显著上升。SRAM 不再只是配角，而是直接影响良率、功耗与系统稳定度的核心组件。

在此趋势下，SRAM 修复显得日益重要，且 已从辅助角色，转变为支撑成长的关键技术。当芯片内部存储器规模扩大，任何微小缺陷都可能造成整颗芯片报废；高效且精准的修复机制，成为控制成本与确保量产稳定性的必要条件。但在大容量与先进制程条件下，传统内存测试方式面临测试时间过长、pattern 冗余与成本上升等问题。

因应此挑战，芯测科技 START™ v5 (荣获 2025 EE Awards Asia 年度最佳 EDA 工具) 以专利化算法与 UDA 模块化架构，让工程师可依制程与产品特性调整测试策略，在提升良率与可靠度的同时，优化测试效率与量产成本。

延伸阅读：

- **从TPU、LPU、NPU 到 CIM：为何先进 AI 计算都离不开 SRAM？**
  https://www.istart-tek.com/2026/01/08/10737/
- **当 TOPS 不再等于效能 AI 算力真正的瓶颈在哪里？**
  https://www.istart-tek.com/2026/01/21/10806/

当内存成为稀缺资源，谁能掌握量产与成本的主控权？

DRAM 与 NAND Flash 进入了新一轮结构性涨价周期，内存已成为直接影响系统成本、交付节奏与供应风险的关键资源。这一轮涨价的背景，并非单纯来自传统终端需求回温，而是 AI 推论、车用电子与边缘计算，对高密度、高可靠度内存的长期消耗正在成为常态。

在此环境下，芯片设计与系统厂更实际面对的问题，是如何在供应受限的前提下，把现有内存资源转化为可量产的产品。特别是在车用与 AI 应用中，内存失效不仅影响良率，更可能牵动整体配置与产品上市时程。

芯测科技的内存测试与修复解决方案，正是因应此产业型态转变的最佳利器。透过完整的 BIST、BISR 与 Repair 架构，客户可在不增加制程成本的前提下，提高可用内存比例，降低报废率，缓解内存单价上升所带来的压力。

同时，芯测以 START™ v5 平台为基础，提供 UDA (User-Defined Algorithm) 机制，让客户能依据实际 AI 计算负载或车用使用情境，客制化测试与修复算法，而非受限于制式模板。这对于长时间运作、高存取频率的应用场景尤为关键。

在内存供应趋于紧张、价格中枢上移的时代，测试与修复已成为影响成本控制与供应确定性的核心能力。透过芯测长年在此领域的耕耘，将可协助客户有效且系统化的管控内存风险，取得最佳解决方案。

立即咨询芯测科技：https://www.istart-tek.com/contact/

辉达 GTC 突显 AI 时代的内存新赛局：从 SRAM 到 MLC NAND 的良率保卫战

辉达 (NVIDIA) GTC 2026 大会甫落幕，全球科技产业的目光再次聚焦于「AI 推论 (Inference) 」。执行长黄仁勋在演讲中明确指出，我们已经进入了「代币工厂 (Token Factory) 」的时代，计算需求在短短两年内暴增了百万倍。

在这场计算风暴中，半导体架构正经历一场前所未有的变革。过去我们关注的是计算核心的算力，但现在内存的表现才真正决定了 AI 代理人 (Agentic AI) 的反应速度和营运成本的表现。

异构推论新趋势：当 GPU 遇上 LPU

在 GTC 相关报导和延伸讨论中，最受瞩目的技术转向莫过于「解构式推论 (Disaggregated Inference) 」。辉达整合了 Groq 团队的技术，推出全新的 Groq LP30 芯片。这背后的逻辑在于解决「内存墙」的瓶颈：Vera Rubin GPU 擅长处理「预填 (Prefill) 」阶段，利用其强大算力处理海量上下文。Groq LPU 则专攻「代币生成 (Decode) 」阶段。它的设计核心是舍弃传统的高带宽内存 (HBM) ，改为搭载庞大的片上 SRAM。

为什么是 SRAM？因为它的延迟极低、速度极快。当 AI 在进行实时对话或生成程序代码时，数据能直接在芯片内部快速流动，不必频繁跳出芯片去存取外部内存。因此，SRAM 的容量与稳定性决定了 AI 芯片的竞争力。

2D NAND 的意外逆袭：MLC 价格飙涨的背后

正当先进制程领域疯狂追求 SRAM 与 HBM 之际，成熟制程市场也爆发了震荡。根据产业最新情资，利基型内存市场正出现「价格翻倍」的奇异现象。

三星、美光等国际大厂为了腾出产能给利润更高的AI应用 (如 300 层以上的 3D NAND) ，正加速淡出低容量、旧制程的 2DNAND 产线。这导致了网通、数字电视与机顶盒等设备必备的MLC NAND 陷入严重断链，2026 年的供应量预计减半，市场缺口高达 3 至 4 成。

此现象反映出，即便在非 AI 核心计算的利基设备中，内存的可靠度与稳定供应依然是企业生存的命脉。当面临 NAND 史无前例的天价，如何确保每一颗出厂内存的完美质量，成为了 IC 设计公司与制造商获利的关键。

芯片良率：AI 时代的隐形护城河

不论是 GTC 大会上备受推崇的 SRAM 核心 LPU，还是市场稀缺的 MLC NAND，都有一个共同的技术挑战：良率与可靠性。

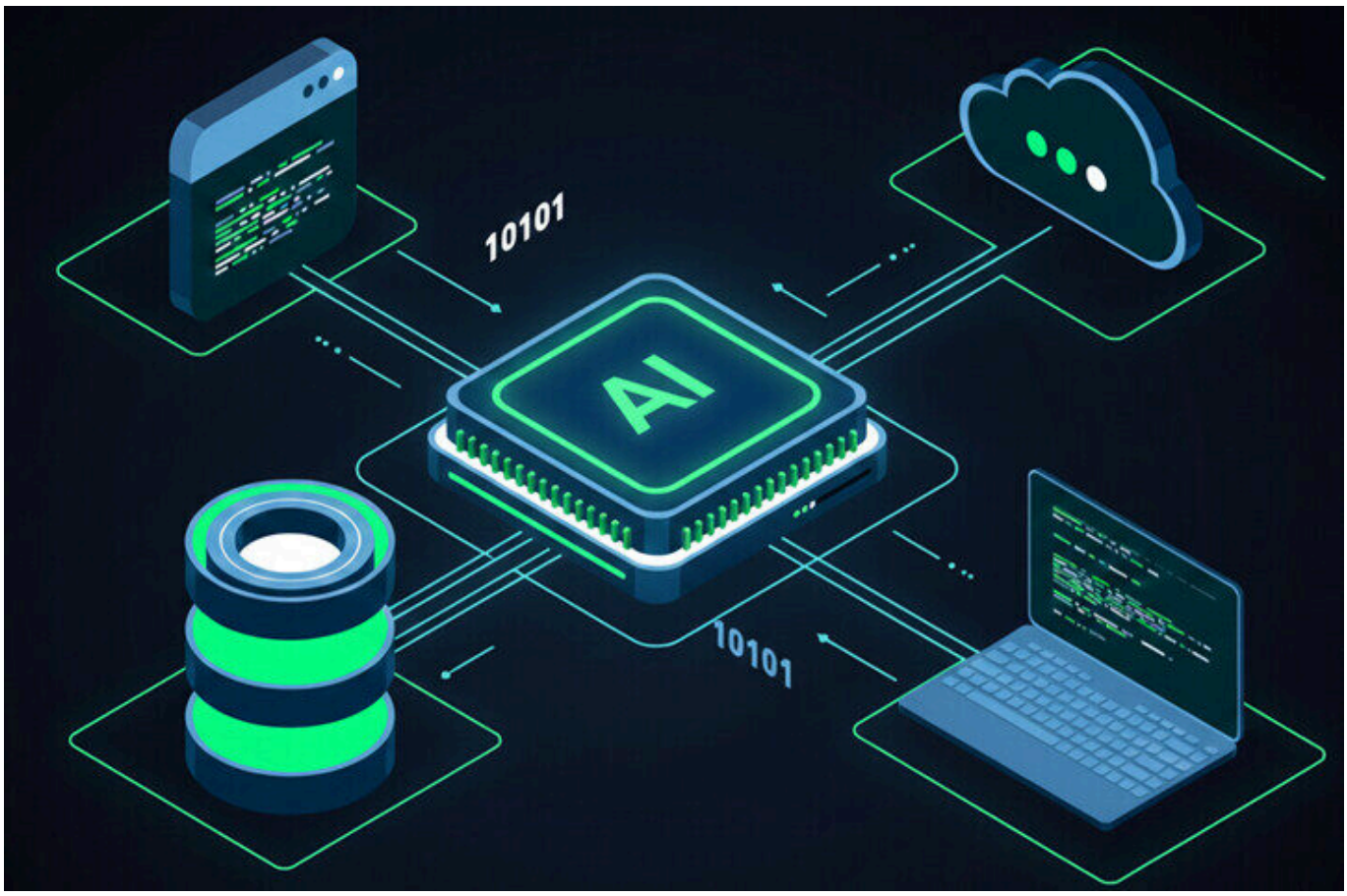
随着先进制程迈向 N3P 甚至是更精细的维度，SRAM 在芯片中所占的面积越来越大，发生故障的机率也随之增加。如果一颗昂贵的 AI 芯片因为内部的 SRAM 损坏而报废，那将是企业巨大的财务损失。这也是为什么投资市场近期高度看好相关测试技术的原因——在追求极致效能的道路上，「测试与修复」已从以往的辅助角色，跃升为守护芯片设计质量的标配。

智能化与高弹性内存测试修复技术 成为获利关键

在这波 AI 推论与内存缺货潮中，芯测科技凭借着深耕多年的内存测试与修复技术，成为 SRAM 良率的守门员。如果说标准算法是基础防护，芯测的用户自定义算法平台 (User-Defined Algorithm, UDA) 就是一座灵活的研发基地。透过图形化接口与独家的 TEC (Testing Elements Change) 技术，工程师无需编写复杂程序代码，即可针对高温、低压等特殊应用场景，自行量身打造测试策略。无论是侦测复杂的 Leakage Defect 还是多样化的内存缺陷 (如 SAF、DRDF) ，UDA 都能实现全面性的防御效果，确保芯片在任何严苛环境下都能稳定运作。

同时，面对先进制程下日益复杂的 SRAM 架构，芯测推出的 MART (MBIST Algorithm Recommendation Tool) 将 AI 技术导入测试流程，克服传统表格查询的繁琐和缺乏弹性的测试选择问题。用户仅需透过简单的互动问答，系统便会根据功耗、面积、良率等条件自动进行 AI 权重加权分析。这不仅大幅降低了决策成本与 DPPM (每百万颗不良品数) ，更让 BIST 算法的挑选从「人工经验」进化为「智慧化精准推荐」，协助客户在激烈的 AI 芯片赛局中夺得上市先机。

先进制程与 AI 服务器架构演进：独立式 NOR Flash 在系统管理中的战略转型



在当前高效能计算 (HPC) 与 AI 服务器的快速演进中，内存在系统架构中的角色正经历变革。市场对数据移动成本、系统功耗与整体运行效率的考虑日益增加，非挥发性内存已从单纯的被动储存组件，也逐步成为参与系统资源分配、安全管理与引导运作的关键环节。在此背景下，嵌入式闪存 (eFlash) 与独立式 NOR Flash 的定位出现了明显的消长。

过去在微控制器 (MCU) 与低功耗边缘计算 SoC 中，开发者倾向于将 eFlash 直接整合于芯片内部，以维持高整合度并降低系统复杂度。然而，当半导体制程节点持续缩小至 28nm 以下，甚至进入 FinFET 先进制程时，eFlash 面临了严峻的结构性限制。由于 eFlash 写入时需要高电压组件与特殊的制程模块，这与先进逻辑制程追求的低电压、超薄闸极氧化层特点产生了物理上的互斥。

强行整合大容量 eFlash 通常需增加约 8–12 道额外光罩 (依制程平台与内存架构而异)，显著提升生产成本，更会面临良率下降与可靠度控制难度增加的挑战。在先进制程中，若持续追求 eFlash 的整合，已不再符合经济效益。

这一技术瓶颈促使系统设计转而采用外挂式储存方案，使得独立式 NOR Flash 在多种高效能应用中重新获得重视。透过 QSPI、OSPI 或其他高速串行接口，SoC 可将韧体、启动程序与关键设定参数储存在外部的专用内存芯片中。这种分工模式使逻辑芯片能专注于高密度计算与 AI 加速，而将非挥发性储存交由具备专业制程的内存厂商负责，提升了整体的制程弹性与供应链成本控制能力。

这种转变在 AI 服务器架构中尤为关键。尽管 AI 系统的主要计算与数据交换依赖于 HBM 或 DDR5 等挥发性内存，但负责系统底层管理的「基板管理控制器」(BMC) 则对 NOR Flash 有着极高的依赖。由于 AI 服务器结构日益复杂，内部包含大量的 GPU、FPGA 与网络交换组件，BMC 必须透过独立式 NOR Flash 来储存核心韧体、执行远程监测与故障诊断。同时为了因应资安威胁，系统通常需要内建安全启动机制 (Root of Trust) ，这类加密签章与安全引导程序亦需储存在高可靠度的 NOR Flash 中。

此外，AI 服务器对可靠度的标准远高于一般消费性电子产品。任何启动错误或韧体损毁都可能导致整个计算丛集停机，产生高额的营运损失。因此数据中心使用的 NOR Flash 在失效率 (DPPM) 控制、耐久度与在高温环境下的数据保存时间方面，必须符合极为严格的验证规范。这种高可靠度需求，使 NOR Flash 从单纯的开机组件，转变为确保系统稳定性与安全性的战略组件。随着 AI 系统导入更频繁的韧体更新与多版本管理机制，NOR Flash 的读写循环寿命与数据完整性验证变得越发重要。

从产业结构来看，这种变化促成了半导体供应链的重新分工。过去 eFlash 是 SoC 厂商的核心整合能力，但在 AI 与先进制程环境下，将储存功能外移至专业内存供货商，反而有助于提升整体效率。逻辑设计公司可专注于计算架构与加速器设计，而内存厂商则在高可靠度市场中建立技术壁垒。总体而言，先进制程下的整合门坎，为外挂式 NOR Flash 创造了新的增长空间。AI 服务器市场持续扩张，推升了高可靠度非挥发性储存的需求，这不仅确立了 NOR Flash 在系统管理架构中不可或缺的地位，也使其成为支撑 AI 时代算力稳定运行的重要幕后功臣。

当 TOPS 不再等于效能 AI 算力真正的瓶颈在哪里？

近几年，AI 芯片的发表会几乎都离不开一个关键词：TOPS。无论是数据中心 GPU、车用 SoC，还是边缘 AI 处理器，算力数字一代比一代高，从数十 TOPS、数百 TOPS，快速跳到上千甚至上万 TOPS。表面上看，AI 算力似乎已经不是问题，但在实际系统设计与应用中，效能瓶颈却仍然频繁出现。这也让工程师开始重新思考一个问题：TOPS 真的是决定 AI 效能的核心指标吗？

TOPS (Tera Operations Per Second) 代表芯片在特定计算精度条件下，每秒可完成的理论计算次数。多数 AI 芯片会以 INT8、INT4 等低精度计算作为标准，因为这类计算最符合推论场景，也最容易堆栈出漂亮的算力数字。然而，TOPS 本身并不等于实际应用效能，它更像是一颗引擎的最大马力值，无法反映整个系统是否能长时间稳定地输出这个效能。

以 NVIDIA 为例，近几代数据中心 GPU 的 AI 算力早已进入万 TOPS 等级。H100、B200 等产品在低精度 AI 计算模式下，理论算力极高，足以支撑大型语言模型与生成式 AI 的推论需求。在终端与边缘市场，NVIDIA Jetson 系列、Qualcomm、MediaTek、Apple 以及 Google 等，也陆续推出具备数十到数百 TOPS 的 NPU SoC，用于影像辨识、语音处理与本地端 AI 推论。从规格表来看，AI 算力似乎已全面到位。

但实务上，AI 推论并不是单纯的计算而已。每一次计算之前，都必须先从内存读取权重与特征数据，计算完成后，再将结果写回内存。当模型规模变大、数据重用频率提高，系统的效能往往受限于数据搬移，而非计算单元本身。这也是为什么在许多应用中，即使芯片标示的 TOPS 很高，实际效能却无法线性成长。

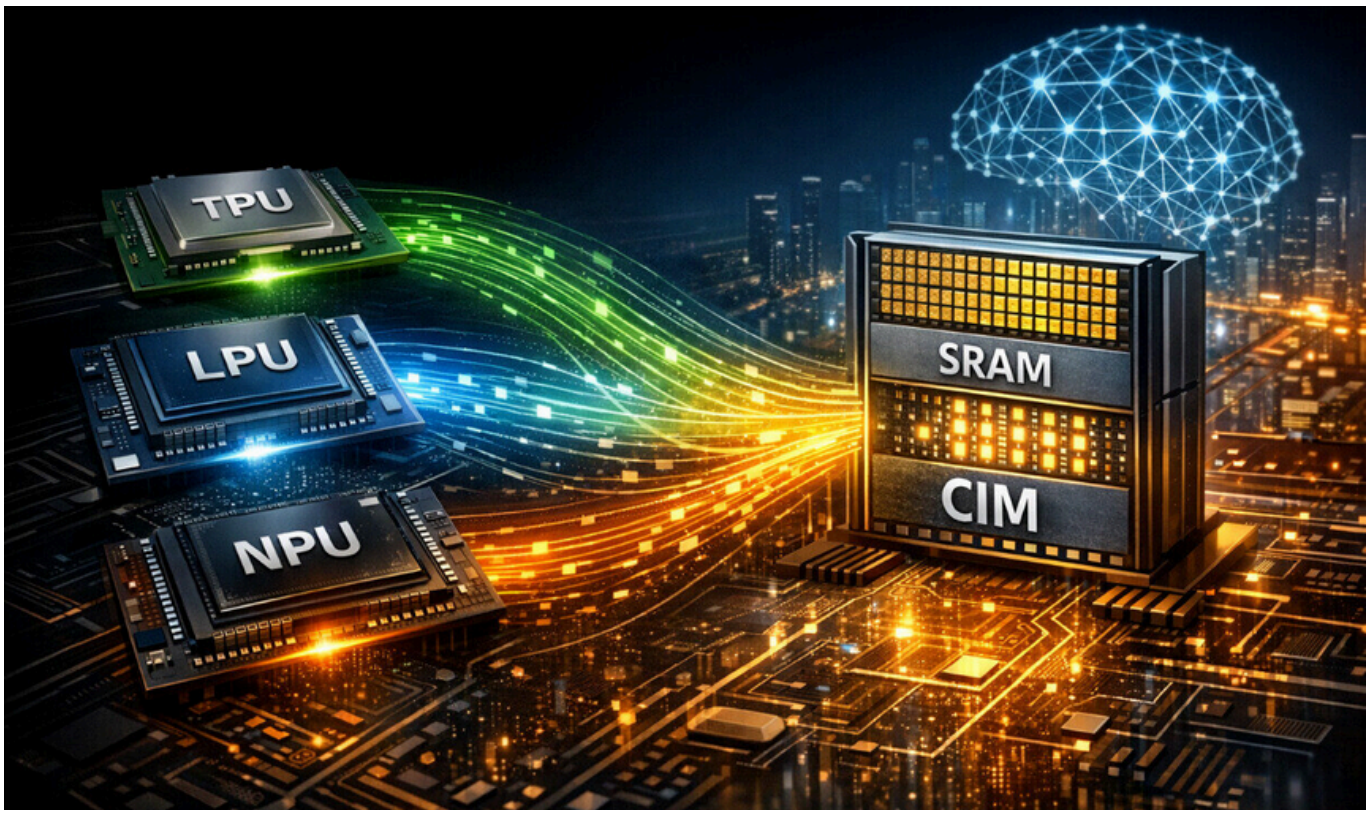
因此我们须回归到内存架构问题。为了支撑高 TOPS 计算，各大 AI 芯片厂纷纷加强内存带宽设计。数据中心 GPU 大量采用 HBM，透过堆栈式封装与超高带宽，缩短计算单元与外部内存之间的距离。而在 SoC 与 NPU 设计中，芯片内部 SRAM 的面积比例则持续攀升，成为不可忽视的关键资源。

SRAM 在 AI 芯片中的角色，并不只是暂存数据这么简单。它负责承接高频、低延迟的数据存取需求，支持权重快取、特征图暂存以及中间计算结果保存。对于卷积神经网络、Transformer 等模型而言，数据的重复存取频率极高，如果每一次都必须回到外部 DRAM，不仅延迟增加，功耗也会快速上升。因此，越来越多 AI 架构选择将关键数据留在片上 SRAM 中，真正发挥 TOPS 所代表的算力。

因为 SRAM 扮演关键角色，其可靠度与测试覆盖率的重要性也随之放大。在高频、长时间运作下，SRAM 容易面临读扰、耦合干扰、老化等问题，一旦发生错误，影响的不只是单一计算，而是整个 AI 推论结果的稳定性。这使得内存测试与修复机制，成为 AI 芯片设计中不可或缺的一环。

从 TOPS 的快速膨胀，到内存架构的持续进化，可以看出 AI 芯片竞争的重心正在转移。算力仍然重要，但真正拉开差距的，往往是谁能让数据流动得更顺畅、内存更可靠。当 TOPS 不再只是宣传数字，而能被完整释放时，AI 芯片的价值，才算真正被发挥出来。

从 TPU、LPU、NPU 到 CIM：为何先进 AI 计算都离不开 SRAM？



近来 NVIDIA 并购 Groq 的消息引发高度讨论，TPU、LPU、NPU 等各类 AI 专用处理器再次成为焦点。无论是云端数据中心、边缘 AI，或车用与工业应用，计算架构正快速从通用 CPU，转向为特定工作负载量身打造的加速器。然而在这些名称各异、定位不同的处理器背后，却有一个共同且关键的核心—— SRAM。

TPU、LPU、NPU 的共通基础

TPU、LPU、NPU 的设计目标虽不相同，但皆高度聚焦于高吞吐量、低延迟与能效比。在实际架构中，大量的权重数据、特征图与中间计算结果，都必须被反复且高速地存取。相较于 DRAM，SRAM 具备低延迟、高带宽与可预测时序的特性，因此成为这类 AI 加速器中最不可或缺的内存组件。

在 TPU 中，SRAM 常被用于大型 on-chip buffer，以支持矩阵计算的数据重用；在强调低延迟推论的 LPU 架构中，SRAM 更是决定实时反应能力的关键；而在各式 NPU 设计里，SRAM 的容量配置、存取并行度与可靠度，往往直接影响整体 AI 效能与功耗表现。

从「内存辅助计算」走向「内存即计算」

当 AI 模型规模持续放大，数据在计算单元与内存之间搬移所消耗的能量，已成为效能与功耗的主要瓶颈。这正是 Computing-In-Memory (CIM) 架构受到高度关注的原因。CIM 的核心概念，是将部分计算直接在内存数组中完成，藉此大幅降低数据搬移次数，突破传统冯纽曼架构的限制。

对于本就大量依赖 SRAM 的 TPU、LPU、NPU 而言，在特定计算场景下导入 SRAM-based CIM，是目前被广泛认为有必要的演进路线之一。透过在 SRAM 内部或周边电路中整合计算能力，AI 加速器得以在维持高可靠度的同时，显著提升能效比，这也使 CIM 成为下一世代 AI 芯片的重要关键技术之一。

CIM 不只是效能 更关乎可靠与安全

CIM 架构所面临的设计挑战，并不仅止于效能。内存本身的变异性、老化效应，以及计算与储存耦合后所带来的测试与验证难度，都让 SRAM 的测试与修复能力变得前所未有的重要。特别是在车用与关键基础设施等高可靠度应用中，CIM 若无完善的测试与修复机制，将难以真正量产与导入。

另一方面，后量子时代的资安需求，也让 AI 与加密计算的结合日益紧密。能否在低功耗条件下，高效执行新一代密码算法，正成为智能装置与车用系统的重要课题。

AI 计算趋势下必备的内存技术

在这样的趋势下，攸关TPU、LPU、NPU 效能与成本的 SRAM 质量，以及 CIM 架构逐渐成为突破数据搬移瓶颈的重要方向。芯测科技长期深耕的 SRAM 测试与修复领域，以及 CIM 架构的研发，将可成为推动技术突破的关键引擎。

芯测致力于提供各类内存之测试与修复专用 EDA 工具与 IP，并针对授权客户，提供涵盖后端流程的一站式设计服务。近年来芯测更积极投入 CIM 架构的研发，重新定义 AI 计算的能效极限。 目前正在开发的 SRAM-based CIM 架构，具备 8-bit 计算精度与极低功耗，并具备良好延展性，可进一步支持 RRAM 架构设计，以满足不同应用场景的需求。

面对量子时代带来的资安挑战，芯测亦率先导入格点密码与 NTT 计算优化技术，透过 CIM 架构加速后量子密码算法，为智能装置与车用系统提供长效且稳固的安全保障。 除了 CIM 架构本身的开发，芯测科技同时也提供 CIM 的测试电路开发环境，让先进内存计算技术不仅能「算得快」，更能「测得准、用得久」。在 TPU、LPU、NPU 持续演进的浪潮中，芯测科技以深厚的 SRAM 技术底蕴，成为推动 CIM 与 AI 芯片走向高效、可靠与节能的重要力量。

当 MRAM 遇上 AI：下一代智能芯片的关键内存革命

随着 AI 芯片算力持续提升，系统效能与功耗的瓶颈正逐渐从计算核心转向内存架构本身。在 AI 推论与训练过程中，大量模型权重与中间数据需频繁于内存与计算单元之间搬移，其能耗与延迟往往高于实际计算本身。为降低数据搬移成本并提升效率，新一代 AI SoC 开始导入 MRAM，利用其兼具高速读写与非挥发特性，使关键数据可长时间常驻芯片内，进一步改善功耗与延迟表现，并支持边缘AI 与低功耗应用需求。

然而，MRAM 的磁性储存机制对制程变异、材料特性与环境条件高度敏感，可能出现写入翻转失败、磁阻值漂移、Retention 衰退或读取边界不稳等失效模式。在高频读写与长时间 AI 工作负载下，这些潜在问题更容易被放大，若缺乏完善的测试与修复机制，将直接影响良率与产品可靠度。因此，在 AI SoC 中导入 MRAM 时，必须建立完整的 BIST 与 BISR 测试修复架构，并进行压力测试、长时间保存测试与跨电压温度条件验证，以确保量产质量并有效控制 DPPM。

芯测科技长期深耕内存测试与修复技术，自主开发的 MRAM BIST IP 已成功导入先进制程 SoC 并完成实际量产应用，具备支持车规等级与高可靠度需求的验证能力。在高频 AI 工作负载情境下，芯测解决方案能提供稳定的测试与修复支持，协助客户兼顾效能、功耗与长期可靠度，并在新一代 AI 内存架构转型中建立良率与量产竞争优势。

观看完整视频

NPU 运作需求下的 SRAM 重要性与测试策略

在 NPU (Neutral Processing Unit) 的实际设计中，SRAM 往往占据芯片中最大面积、消耗最多功耗，同时也是最容易影响可靠度的关键组件。为了支持高并行度与高度数据重用的AI 推论行为，NPU 会在计算单元周围配置大量 on-chip SRAM，作为权重缓冲区、特征图暂存与局部数据存取空间，确保数据能就地取用并高速周转。

然而，这样的架构也使 SRAM 长时间处于高频率、重复存取的极端使用情境。读扰 (Read Disturb)、耦合干扰、弱写入、感测边界不稳，以及数据保留 (Retention) 相关失效，在 NPU 中更容易被放大，成为影响良率与 DPPM 的主要风险来源。这也使得 NPU 的效能瓶颈，往往不在于算力，而是在内存子系统的设计与测试质量。

本集从 NPU 的实际运作需求出发，解析不同 SRAM 使用情境 (如权重缓冲、特征图暂存、累加结果与待命区域) 所对应的潜在失效风险，并说明为何 NPU 的 SRAM 测试不能仅依赖单一、静态的测试算法，而必须导入可依情境调整的测试与修复策略，才能有效降低量产风险。

针对此挑战，芯测科技以 START™v5 平台为核心，结合 UDA (User-Defined Algorithm) 的图形化与模块化设计，让工程师可依不同 SRAM 结构与使用行为，自行组合最合适的 MBIST 测试流程。同时搭配 MART (MBIST 算法推荐工具)，将既有 SRAM 测试经验系统化，协助在多重限制条件下快速选择适当的测试算法组合，降低决策成本，避免因测试策略不当而影响量产。

透过 UDA 与 MART 的整合应用，SRAM 测试不再只是形式上的验证流程，而能真正贴近 NPU 的实际 AI 工作负载，在兼顾可靠度的同时，提升良率、控制 DPPM，并确保量产效率与产品竞争力。

观看完整视频

LPU 使用 SRAM 的原因与测试关键

随着边缘 AI 与低功耗 AI 推论应用快速成长，LPU (Language Processing Unit)在系统架构设计上，普遍大量采用 SRAM 作为主要 on-chip memory。其关键原因不仅在于速度，而是因为 LPU 推论行为对内存存取延迟、稳定性与可预期性 有极为严苛的要求。本集将从 AI 推论流程出发，解析为何在影像辨识、语音触发与传感器数据分析等实时应用中，内存行为往往比计算效能更容易成为系统瓶颈。

内容将深入比较 SRAM 与 DRAM 在存取延迟、refresh 机制、功耗行为与系统稳定性 上的差异，说明 DRAM 在 LPU 架构中可能带来的不确定性风险，以及 SRAM 在低电压、高频存取环境下的关键价值。同时也说明，在现代 LPU SoC 中，大量、异质化的 SRAM bank 如何显著提升内存测试与验证的复杂度。

本集进一步探讨传统 SRAM 测试方法与实际 LPU 推论存取情境之间的落差，并说明仅以基本读写测试，为何无法有效反映低电压、实际负载下的潜在风险。最后，将介绍芯测科技如何透过 START™ v5 平台，结合 UDA (User-Defined Algorithm) 与 MART (MBIST Algorithm Recommendation Tool) ，协助设计团队依 LPU 应用特性，建立更贴近实际推论行为的 MBIST 测试策略，兼顾可靠度、良率与量产效率。

观看完整视频